International Document Image Processing Summer School 2013

3-7 June 2013

Fourni - Greece

Proceedings

INDEX

About the Summer school

The first Summer School on Document Image Processing tried to provide both an objective overview and an in-depth analysis of the state-of-the-art of this research field and its current open issues. The courses were delivered by world-renowned researchers in the field, covering both theoretical and practical aspects of real problems of Document Image Processing, as well as examples of successful applications.

The participants had the opportunity to present the results of their scientific research, and interact with their colleagues in a friendly and constructive atmosphere.

Historical document retrieval for the Source Corpus Project

Fryni Kakoyianni-Doa and Eleni Tziafa

University of Cyprus, Department of French Studies and Modern Languages, 12 Aglantzias P.O. Box 20537 2108, Nicosia, Cyprus
frynidoa@ucy.ac.cy; tziafa.eleni@ucy.ac.cy

ABSTRACT

The Source Corpus Project is about the Searchable Online French-Greek parallel corpus for the University of Cyprus (SOURCe), which aims to serve the needs of students of French as a foreign language and also to facilitate future linguistic research. This project is led by Fryni Kakoyianni-Doa and is fully funded by the University of Cyprus. We included different registers[1], so that students may compare the results and the use of each word or phrase in different contexts (e.g., literature, scientific, official, technical and political discourse). Since the corpus is designed to be open and freely available to all language instructors and learners, and to be redistributed in toto, the texts included had to be copyright free; hence, they are dating back to the early 20th century or earlier. Due to the phenomenon of diglossia in Greek (katharevousa and dimotiki), and the polytonic system, these texts need a special optical character recognition.

Keywords: Historical Document Retrieval, OCR, parallel corpus, historical corpus, diachronic corpus, Source Corpus, French & Greek language

1. INTRODUCTION

The Source Project focuses on the construction of a parallel corpus, the composition, annotation, encoding and availability of which are meant to serve the needs of students of French as a foreign language and also to facilitate future linguistic research[2]. This project is led by Fryni Kakoyianni-Doa and is fully funded by the University of Cyprus.

Parallel corpora consist of original and translated texts. A parallel corpus, as defined in Ref. 3 in EAGLES typology, is "a collection of texts, each of which is translated into one or more other languages than the original". The corpus is aligned in the sentence level. Based on the firm belief that the days of pencil-and-paper teaching, and not only research[4], are numbered, we decided to proceed to a "data driven learning" approach[5] and the building of an electronic resource for a less resourced language, such as the Greek language. New tools are critical to creating a dynamic and engaging learning environment. Due to lack of such tools, our strategy has been to construct a parallel corpus with a global language (French) and to exploit the resources that already exist for the

global language in constructing analogous resources for the minority language (Greek).

The goal of the project is to provide high-quality, online content, on general education subjects, to students, and also training in a cost-effective manner, as an Open Educational Resource, aiming to be accessible, adaptable, free, high-quality, empowering and relevant.

2. THE SOURCE CORPUS

2.1 The Source Corpus vs Similar Parallel Corpora

The main differences of the Source Corpus, compared to similar projects, such as Linguee (http://www.linguee.com/) or Glosbe (http://en.glosbe.com/), are presented in Table 1:

Table 1. The Source Corpus vs other searchable online corpora (e.g. Linguee, Glosbe)

The Source Corpus	Linguee, Glosbe
Texts available *in toto*, fully downloadable	Parallel sentences only, available for copy paste, but shuffled and out of context
Parallel texts in French and Greek	Linguee: absence of Greek language Glosbe: Limited representation of Greek language, unknown size of corpus
Carefully selected texts, appropriate for use in classroom, according to the design principles of the corpus	No design principles, only collection of all the available parallel texts, mostly from already available parallel corpora (OPUS project) Unappropriate texts for use in classroom included, due to the Opensubs corpus, which includes subtitles from adult movies
Research par register or in specific texts	Research in total corpus only
Literature included	Absence of literature
Non commercial	Commercial (advertisements)
Created by a scientific team, including corpus researchers, experienced teachers and IT experts	Unknown creators, no citations

2.2 Design Principles

Another important issue is choosing appropriate topics, texts and material to be included in the corpus. Most parallel corpora are not register-diversified; nevertheless, our objective is to include at least five different registers[2], so that students may compare the results and the use of each word or phrase in different contexts (e.g., literature, scientific, official, technical and political discourse). In general, the corpus comprises of a fiction and a non-fiction part.

The Source Corpus was compiled with teaching primarily in mind, and also in order to extract translation units from authentic data. As it is intended to be used as teaching material in the classroom, we manually selected the content, in order to be appropriate for learners. It is a diachronic corpus, as the time period covers at least six centuries, from the 15th to 21st century. The texts are copyright free texts, parallel, for the moment in French and Greek language. Research, including comparison between its registers, has already been conducted on a part of the corpus (a sample of 1 million words), while the online searchable corpus includes currently more than 40 million words.

The main objective of this corpus is to provide a "transparent" set of data to teachers and researchers, suitable for linguistic study and research, easy to use, in the form of an open educational resource. The recognition of historical documents is essential for the content exploitation of valuable historical works that bring closer the Greek and French culture.

3. TEXTS INCLUDED IN SOURCE CORPUS

3.1 Collection of Texts

In accordance to design principles and committed to the notion of open source tools and resources, we started searching for copyright free texts, in order to avoid or reduce risks incurred in possible violations of intellectual property rights (IPR) or basic ethical rules. Moreover, these texts have to be redistributable, since it is essential for researchers to have their own corpus to experiment on, instead of using online only searchable corpora.

For this reason, as a first step, we turned to already existing parallel corpora and available parallel texts in French and Greek language (e.g. DGT Multilingual Translation Memory of the Acquis Communautaire[6]), French and Greek aligned texts from Opus project[7] (e.g. ECB, EUconst, EUROPARL, OpenSubs, which contain French and Greek texts) and from the WIT3[8] (Web Inventory of Transcribed and Translated Talks at TED conferences).

Moreover, in the second phase of the project, we had aligned (using the open source alignment tool Lf aligner) French and Greek literature works from Project Gutenberg or Wikisource, (http://fr.wikisource.org, http://el.wikisource.org). We also added new texts, such as technical texts, manuals (e.g. Linux, Php).

Finally, in the third phase, we searched and collected texts (in the literature domain) from online digital libraries (http://gallica.bnf.fr,

http://books.google.com, http://anemi.lib.uoc.gr/, http://argolikivivliothiki.gr, http://pergamos.lib.uoa.gr/, http://www.snhell.gr, http://www.iaen.gr, http://xantho.lis.upatras.gr, http://openarchives.gr, http://abu.cnam.fr, https://lekythos.library.ucy.ac.cy, etc.) and physical libraries in Greece and Cyprus, after consulting the Index Translationum and bibliographical lists of translated works in both languages. We also scanned available and copyright free books. Finally, we asked for donations by publishing houses and translators.

3.2 Problems Encountered, Related to OCR Issues

In this type of research, which focuses on parallel corpora, the difficulty is twofold: the scarcity of parallel corpora and texts with their translations, and the scarcity of Greek texts in electronic form, and the even more rare representation of Greek language in parallel corpora.

The already existing parallel French-Greek corpora had only minor typographic errors due to OCR. Moreover, most of the texts from the 20th and 21st centuries did not cause major problems as regards scanning. However, this was not the case with older, historical documents, both in French and Greek language.

There are three types of texts, as regards their form, that had to be processed and:

- Texts already in electronic text form, such as those in Project Gutenberg or Wikisource
- Texts available scanned as images, such as those from Gallica for French or Anemi and 66 other Greek digital libraries searched through Open Archives for Greek (http://openarchives.gr)
- Texts available as physical books in libraries or donated by publishing houses that had to be scanned and processed to text.

The texts deriving from Project Gutenberg were clear, manually corrected (by Sophia Canoni) electronic texts; nevertheless, probably for practical reasons and readability, the tonic system has been changed from polytonic to monotonic. In a corpus designed for teaching, it would be useful to include also the original polytonic texts which are available in the public domain, in order to study and search different forms of Greek language.

ΖΟΥΡΝΤΑΙΝ
Ναι.

Ο ΚΑΘΗΓΗΤΗΣ ΤΗΣ ΞΙΦΟΜΑΣΚΙΑΣ
Και εξ αυτού εμφαίνεται πόσο πρέπει να μας υπολήπτεται εμάς το κράτος, και κατα πόσον η γνώσις της ξιφομαχίας υπερέχει από όλας τας αλλας ανωφελείς γνώσεις, καθώς, παραδείγματος χαριν, από το χορο, από τη μουσικη, απο...

Ο ΧΟΡΟΔΙΔΑΣΚΑΛΟΣ
Σας παρακαλω, κύριε ξιφομάχε! πρέπει να μιλήτε με σεβασμό για το χορο.

Ο ΜΟΥΣΙΚΟΔΙΔΑΣΚΑΛΟΣ
Μαθετε να μιλήτε καλύτερα για τη μουσικη.

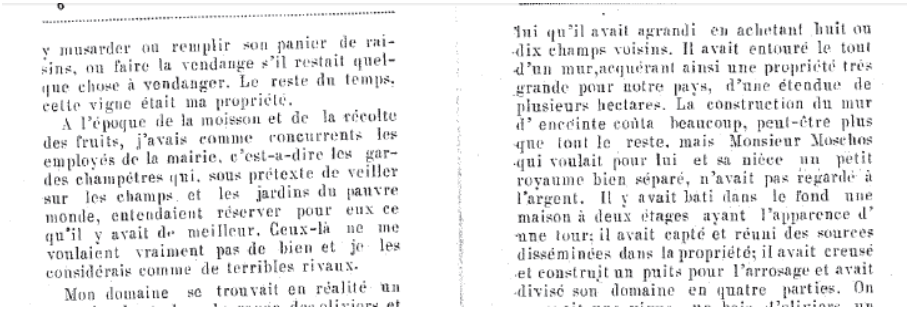
Ο ΚΑΘΗΓΗΤΗΣ ΤΗΣ ΞΙΦΟΜΑΣΚΙΑΣ
Είσθε πολύ αστειοι, αν θέλετε να συγκρίνετε την τέχνη σας με τη δικη μου.

Ο ΜΟΥΣΙΚΟΔΙΔΑΣΚΑΛΟΣ
Τι ιδέα που εχει για τον εαυτό του!

Ο ΧΟΡΟΔΙΔΑΣΚΑΛΟΣ
Πολύ κωμικός μου φαίνεσθε με τον ψευτοθώρακά σας.

Figure 1. An excerpt from the Greek translation of *The Imaginary Patient* ("Le bourgeois gentilhomme") by Moliere, as provided by Project Gutenberg in *.txt format

The texts available from digital libraries and also the text scanned from physical libraries are mostly images in *.pdf, *.tiff or *.jpg formats, usually in 300dpi resolution:



y musarder ou remplir son panier de raisins, ou faire la vendange s'il restait quelque chose à vendanger. Le reste du temps, cette vigne était ma propriété.

A l'époque de la moisson et de la récolte des fruits, j'avais comme concurrents les employés de la mairie, c'est-à-dire les gardes champêtres qui, sous prétexte de veiller sur les champs et les jardins du pauvre monde, entendaient réserver pour eux ce qu'il y avait de meilleur. Ceux-là ne me voulaient vraiment pas de bien et je les considérais comme de terribles rivaux.

Mon domaine se trouvait en réalité un

lui qu'il avait agrandi en achetant huit ou dix champs voisins. Il avait entouré le tout d'un mur,acquérant ainsi une propriété très grande pour notre pays, d'une étendue de plusieurs hectares. La construction du mur d'enceinte coûta beaucoup, peut-être plus que tout le reste, mais Monsieur Moschos qui voulait pour lui et sa nièce un petit royaume bien séparé, n'avait pas regardé à l'argent. Il y avait bati dans le fond une maison à deux étages ayant l'apparence d'une tour; il avait capté et réuni des sources disséminées dans la propriété; il avait creusé et construit un puits pour l'arrosage et avait divisé son domaine en quatre parties. On

Figure 2. An excerpt from the French translation of the Dream on Waves by Alexandros Papadiamantis, as provided by the Library of the University of Crete in pdf format

Texts from physical libraries were selected to be as free of lines, marks or smudges as possible and they were scanned in 300dpi resolution. Most of the books scanned had distortions such as colored pages (yellow to reddish) or shrinking due to humidity. The following steps were followed in the processing of texts from physical libraries:

- We used the original hard copies, the cleanest version possibly
- The scanning resolution was 300dpi, in order to achieve as rapid results as possible
- The *.tiff file format was selected, so that no image information (pixels) would be lost

– The image was scanned as grayscale, with increased contrast and density

– Pages were de-skewed, so that word lines were horizontal

3.3 OCR Software Used

For the OCR process we used ABBYY FineReader 11 Professional, mainly because the texts very often came in columns, and the program had the ability to decolumnize the texts, and also because of the training ability, which can be shared among different users of the same software.

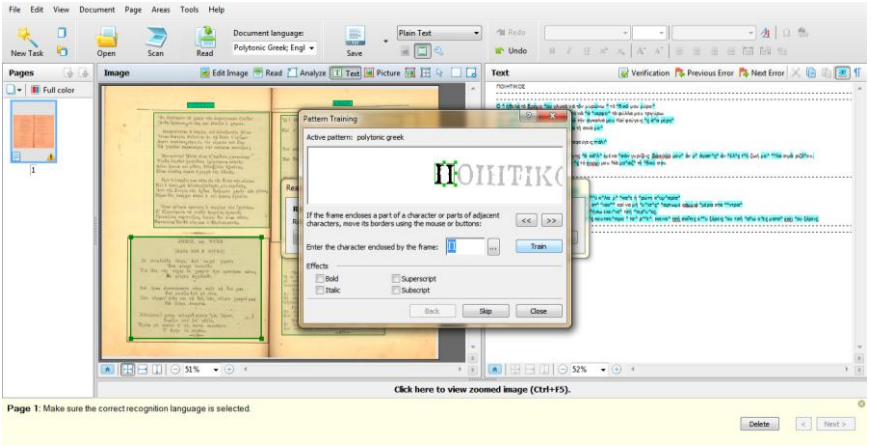


Figure 3: Training of ABBYY FineReader 11

We also tried Tesseract (https://code.google.com/p/tesseract-ocr), but the results were much poorer, and training required more time. Therefore, we conducted extensive training of ABBYY FineReader 11 and formed a user pattern, but what we seek is advice and support by computer specialists for choosing suitable OCR methods but also for creating the final, published and distributed corpus.

4. CONCLUSION AND FUTURE WORK

As pointed out in Ref. 9, “to this date, polytonic, or ancient, Greek has not joined the revolution in reading and textual study brought about by libraries' participation in large-scale optical character recognition (OCR) projects like Google Books. There are good reasons for this. Ancient Greek comprises vowel accents and breathing marks that can easily confound a OCR engine, and over the years a great variety of font faces have been used to represent the language. Greek ought not to be left behind in this, not only because many books were published primarily in Greek, but also, perhaps more importantly, because books in modern Western languages have, since the invention of the printing press, drawn on ancient Greek as an intellectual heritage. They quote Plato, Galen, Aeschylus and the Church Fathers to explore modern ideas. If OCR processes

render these quotations as indecipherable misreadings, this particular web of meaning, tracing across languages and time, remains inaccessible".

As regards texts derived from project Gutenberg, we intend to include also the polytonic form of the texts. We also plan to enrich the corpus as much as possible with texts from digital and physical libraries.

We are interested in working with an open source OCR system, in order to be able to share not only the texts included in the corpus but also the training data set and our dictionaries, in order to serve as an aid for future document image processing of new texts.

Copyright issues were a major limitation in this project as it is aimed for educational purposes, but it was this limitation that led us to a form a corpus with diachronic texts, with the most influential books. With this effort, and exploring all possible ways to recognize and include historical Greek texts in a corpus used for language learning, we hope to contribute to the construction of valuable resources for a less resourced language.

ACKNOWLEDGMENTS

We would like to thank our IT technicians Stefanos Antaris and Xenia Christodoulou. We gratefully acknowledge funding by the University of Cyprus.

REFERENCES

1. D. Biber, "Representativeness in corpus design", *Literary and Linguistic Computing* **8/4**, pp. 243-257, 1993.
2. F. Kakoyianni-Doa, Tziafa, E., "Source: Building a Searchable Online French Greek Parallel Corpus for the University of Cyprus", in *Revista Nebrija de Lingüística Aplicada* **11** *(número especial)*, (forthcoming).
3. J. Sinclair, C. Ball, "Preliminary Recommendations on Text Typology", *EAGLES Documents EAG-TCWG-TTYP/P*, 1996. <http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html>
4. T. Váradi, S. Krauwer, P. Wittenburg, M. Wynne, and K. Koskenniemi, "CLARIN: Common Language Resources and Technology Infrastructure", *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pp. 1244-1248, 2008. <http://www.lrec-conf.org/proceedings/lrec2008/pdf/317_paper.pdf>
5. T. Johns, "Should you be persuaded: two examples of data-driven learning", in *Classroom Concordancing*, T. Johns et P. King, dir., *English Language Research Journal* **4**, pp. 1-16, 1991.
6. R. Steinberger, A. Eisele, S. Klocek, S. Pilos, and P. Schlüter, "DGT-TM: A freely Available Translation Memory in 22 Languages", in *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*, 2012 <http://langtech.jrc.ec.europa.eu/Documents/2012_LREC_DGT-TM_Final.pdf>.

7. J. Tiedemann, “Parallel Data, Tools and Interfaces in OPUS”, in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2012 <http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf>
8. M. Cettolo, C. Girardi, and M. Federico, “WIT3: Web Inventory of Transcribed and Translated Talks”, in *Proceedings of the 16th EAMT conference*, pp. 261-268, Trento, Italy, 2012 <https://wit3.fbk.eu/papers/WIT3-EAMT2012.pdf>
9. B. Robertson, Large-Scale Polytonic Greek OCR - e-Humanities Home, 2012 <http://www.e-humanities.net/assets/seminar/2012/Robertson.pdf>

Digitisation of the Complete Works of E. Kriaras in the Portal for the Greek Language

John N. Kazazis, Rania Voskaki,
Athanasia Margoni, Christos Andras

Centre for the Greek Language
jkazazis@komvos.edu.gr, rvoskaki@hotmail.com
ath.mrg@gmail.com, andraschris@gmail.com

ABSTRACT

In this paper, the digitisation of the *Complete Works of Emmanouel Kriaras*[1] in the Portal for the Greek Language[2] is presented. It is a project undertaken by the Centre for the Greek Language (CGL) in order to allow both the specialised researcher and the average reader to gain easy access to his works. In this framework we have digitised the entirety of his scientific work and have accompanied them by indices of terms (i.e. key-words of each work), which are electronically linked to the pages of the work they refer to. The unit created in the Portal for the Greek Language contains, at the time being, 323 works and more than 163,000 indices of terms.

Keywords: digitisation of scientific works, retrieval of indices of terms

1. INTRODUCTION

The main aim of the CGL is the overall support and promotion of the Greek language in Greece and abroad. For this purpose, it has implemented the Portal for the Greek language[1] comprising electronic tools (i.e. online dictionaries and text corpora) and language material (i.e. anthologies and studies) addressed to researchers, university students, teachers, pupils and everyone interested in the Greek language within and outside Greece. Therefore, it attempts to cover the Greek language both diachronically and synchronically (Ancient Greek, Medieval Greek, and Modern Greek) by supporting and disseminating the Greek language in the digital age. Under this scope the digitisation and the web uploading of an extended electronic corpus, such as the entire works of Kriaras, covering both Medieval and Modern Greek was imperative.

For the implementation of the project, we were mainly based on the Kriaras archive (all in hardcopy) and on three anthologies of his works[4,5,6] (hardcopies as well). The retrieved archive material was gradually digitised and divided in the following thematic units:

1. Dictionary of Medieval Vulgar Greek Literature (17 volumes).
2. Medieval Studies: Scientific studies by E. Kriaras on texts of medieval literature (13 works).
3. On Language: Studies on the Greek language and the Greek Language Problem (27 works).
4. Correspondence - Autobiographical Works - Other Documents (15 works):
   a. books with E. Kriaras's correspondence,
   b. autobiographical works, and
   c. rare written documents.
5. Monographs-Book reviews (13 works): other authors' attempts to compile lists of the works of E. Kriaras, and various publications and reviews on his work.
6. Journals (238 works): articles in scholarly journals by E. Kriaras, as well as by other prominent scholars regarding his work.

Also, audio-visual material has been uploaded: videos with interviews, speeches, and opening speeches by E. Kriaras; 7 in total at the time being.

The contribution of the present work consists of:

- The digitisation of scientific works of E. Kriaras.
- The categorisation of the uploaded material in thematic units.
- The representation of the digitised material in a human-readable format.
- The creation of an electronic index, allowing users of the Greek Portal the automatic retrieval of indices of terms linked to the wider context they belong to.

2. DIGITISED DATA AND METHODOLOGY

The first step was to scan the archive material and make them accessible on the Greek Portal. More precisely, the online available material consists of the following number of data:

Table 1. Quantity of digitised material available online.

Thematic Units	**Number of Works**	**Pages in total**	**Page average per Work**
Dictionary of Medieval Vulgar Greek Literature	17	6,852	403
Medieval Studies	13	3,451	265
On language	27	7,948	294
Correspondence, Autobiographical works, Other documents	15	3,556	237
Monographs, Book Reviews	13	1,380	106
Journals	238	2,182	9
Total	323	25,369	78.541

The above mentioned data present the following characteristics:

- the resolution is medium, 300 dpi,
- the pages, book size, are scanned one at the time,
- evenly lighted images,
- the file format is png,
- most of the scanned pages are deskewed, with the exception of certain images slightly skewed,
- most of the material is in good quality, except from a few old printing of bad quality.

We opted for 300 dpi resolution in order to have low computational cost both in terms of space and time. It is worth mentioning that, in the scanner we used, 300 dpi need 13 seconds, 400 dpi require 18 seconds and 600 dpi necessitate 50 seconds per page respectively. In addition, by choosing a higher resolution analysis we would come up with a considerably heavier material that would be unmanageable while uploading and using it in the Greek Portal.

In parallel, the indices of terms were recorded in a database, in order to study the variables of Kriaras's archive that should also be taken into consideration (i.e. page number, notes, citations). In particular, further information relating to the indices of terms had to be noted down (i.e. type of document, author, and publisher). Therefore, a relational database was designed in a Microsoft SQL Server 2008. It is worth mentioning that during the project, the digitised material turned out to be diverse, non-homogeneous, and differentiated and did not follow a typical model. There were many differences between publishers, between different editions of the same publisher and, of course, between the various types of document (books, journals, hand-written texts, etc.). For this reason, the database was gradually enriched by additional fields in order to achieve the best possible representation of the digitised material.

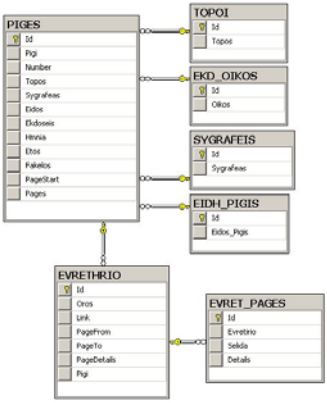


Figure 1. SQL database schematic.

The basic table-entity is the table entitled “PIGES”, which comprises all the different units of the complete works of E. Kriaras (i.e. the information of a book, of a dissertation, of an article, etc.).

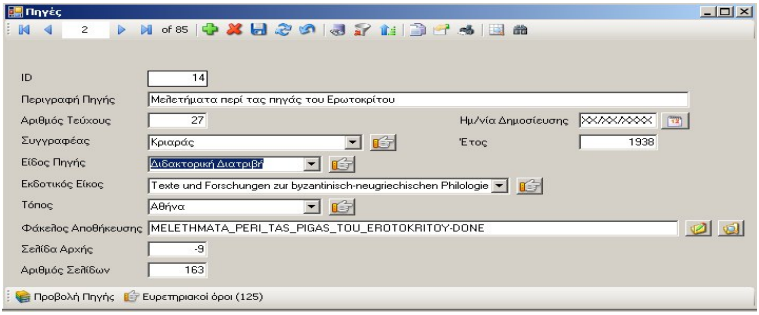


Figure 2. Book entry example in the SQL database.

Every source belongs to and is categorised in a type of document in the Table named “EIDI_PIGIS”.

Figure 3. Types of document sample in the SQL database.

We also recorded the publisher in the Table entitled “EKD_OIKOS” in every entry.

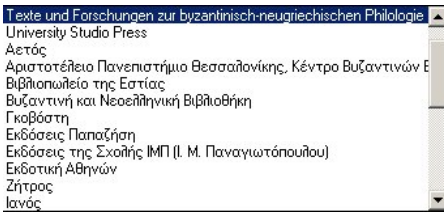


Figure 4. List of Publishers example in the SQL database.

The location of publication was also a category of information entered into the data base, in the Table named “TOPOI”.

Figure 5. Location of publication example in the SQL database.

Finally, the author of the work is recorded in the Table entitled “SYGRAFEIS”. This category is necessary for works that were authored by someone other than Kriaras, e.g. the author of a book review or article.

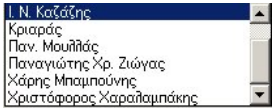
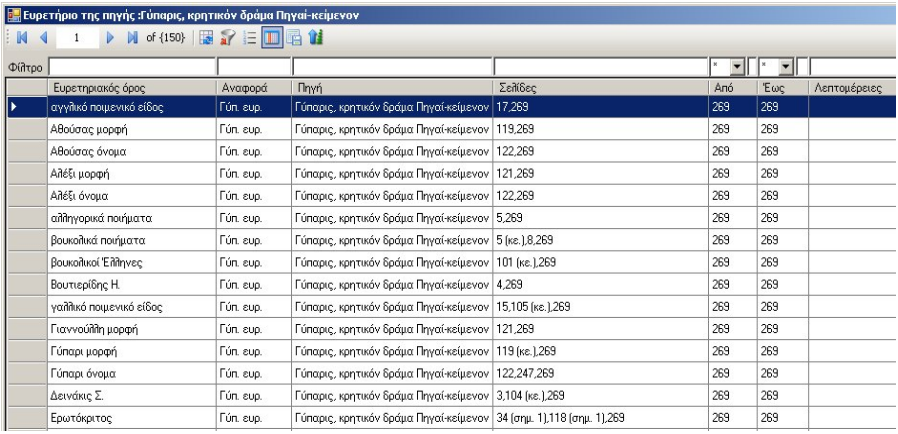


Figure 6. List of authors’ example in the SQL database.

Every source has many index terms recorder in the Table named “EVRETHRIO”. An index term is any word or phrase that the end user could look up in a search. Moreover, every term links to a specific part or page of the source, where it is being used. The following image shows an excerpt of the index terms of one source.



Ευρετηριακός όρος	Αναφορά	Πηγή	Σελίδες	Από	Έως	Λεπτομέρειες
αγγλικό ποιμενικό είδος	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	17,269	269	269	
Αθούσας μορφή	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	119,269	269	269	
Αθούσας όνομα	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	122,269	269	269	
Αλέξι μορφή	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	121,269	269	269	
Αλέξι όνομα	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	122,269	269	269	
αλληγορικά ποιήματα	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	5,269	269	269	
βουκολικά ποιήματα	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	5 (κε.),8,269	269	269	
βουκολικοί Έλληνες	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	101 (κε.),269	269	269	
Βουτιερίδης Η.	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	4,269	269	269	
γαλλικό ποιμενικό είδος	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	15,105 (κε.),269	269	269	
Γιαννούλη μορφή	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	121,269	269	269	
Γύπαρι μορφή	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	119 (κε.),269	269	269	
Γύπαρι όνομα	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	122,247,269	269	269	
Δεινάκις Σ.	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	3,104 (κε.),269	269	269	
Ερωτόκριτος	Γύπ. ευρ.	Γύπαρις, κρητικόν δράμα Πηγαί-κείμενον	34 (σημ. 1),118 (σημ. 1),269	269	269	

Figure 7. Excerpt of the index terms of one source in the SQL database.

In the detailed entry form where we entered the information for each term, we link it to the source and the page numbers it refers to in the Table named “EVRET_PAGES”.

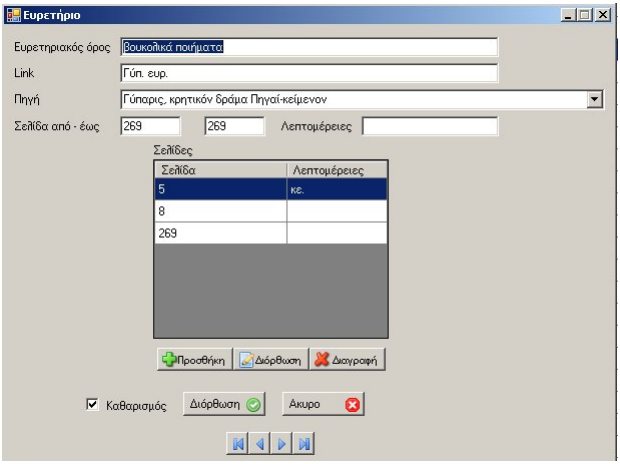


Figure 8. Index term example in the SQL database.

Finally, we automatically transferred the data sample in an adequately created database at the Greek Portal and further recording of new entries is directly inserted in it. In the next figure we cite a term example:

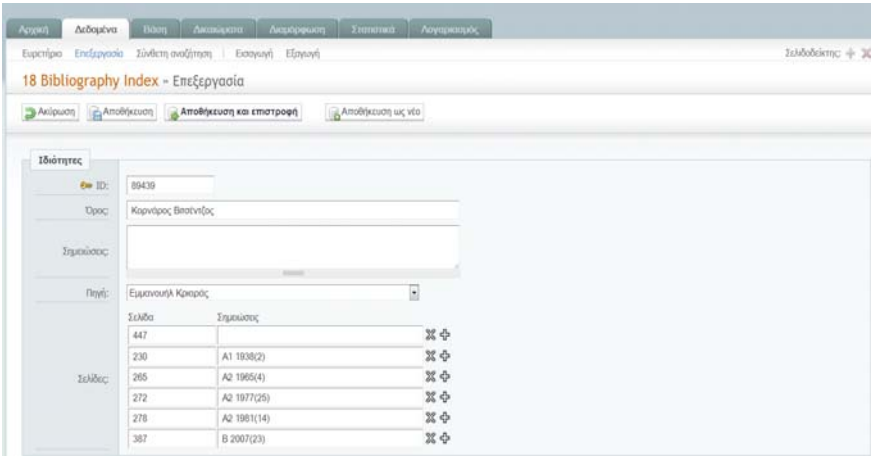


Figure 9. Term example in the Greek Portal's database.

In addition, a terms search machine was created in the Greek Portal. It allows:

a. the retrieval of a given term out of the entire online material (Fig. 10),
b. the retrieval of the terms included in a specific work (Fig. 11) and
c. the link to the page and/or pages of the work that they refer to.

Figure 10. General term research.

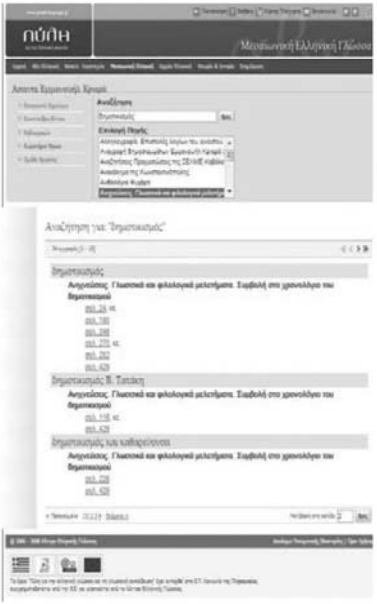


Figure 11. Specific term research.

3. EVALUATION AND PERSPECTIVES

By following the above mentioned fundamental methodology, in terms of digitisation, we achieved to make accessible a considerable number of works. It is worth mentioning that they constitute the majority of the existing works of and on Kriaras. Approximately 15 books are kept in abeyance (although they have already been digitised) due to editors' rights, as well as new editions that are to be published in 2013, articles in scholarly journals, and audio-visual material (interviews).

Moreover, the statistics offered by google analytics, regarding the website's traffic, prove the utility of the unit created in the Greek Portal along with the interest shown by the users. In the past two years

70,000 users have visited the specific unit; 56.40% were returning visitors, while 43.60% were new visitors.

The improvement of the image processing technique, the qualitative conversion of image files to machine-readable files and the optimisation of seeking performances are among the variety of perspectives that we will endeavour to fulfil through future research. In addition, the integration of the *Dictionary of Medieval Vulgar Greek Literature* along with Greek Portal's online dictionaries is our priority. The online version of the Epitome of the Kriaras Dictionary (two volumes) [3] could therefore be enriched.

Currently, we are testing the results obtained by an OCR conversion. Nevertheless, the medium analysis (300 dpi) of the digitised material causes certain problems in the recognition results. They will have to be evaluated by using concrete evaluation measures such as recall and precision in order to end up with immediate and solid results.

4. CONCLUSION

In this paper, we presented the digitisation of the *Complete Works of Emmanouel Kriaras*, available in the Portal for the Greek Language. The aim was to make the entirety of Kriaras scientific work accessible both to specialised researchers and to average readers. For this purpose we digitised his archive (i.e. books, articles, documents). The Greek Portal contains, at the time being, 7 videos and 323 works, accompanied by more than 163,000 indices of terms.

REFERENCES

1. The Complete Works of Emmanouel Kriaras:

   www.greek-language.gr/greekLang/medieval_greek/em_kriaras/index2.html

2. Portal for the Greek Language: www.greek-language.gr

3. Epitome of the Kriaras Dictionary:

   http://www.greek-language.gr/greekLang/medieval_greek/kriaras/index.html

4. P. Ch. Ziogas, *Emmanouel Kriaras*, Nea Ekdosi, Thessaloniki [in Greek], 2008.

A. A. Lampraki-Paganou, *The Works of Emmanouel Kriaras*, Centre for the Cretan Literature, Heraklion [in Greek], 2001.

5. P. D. Mastrodimitris, *The Works of Emmanouel Kriaras. 70 years*, Domos, Athens [in Greek], 2000.

Combining multiple features based on graphometry for writer identification as part of Forensic Handwriting Analysis

Aline Maria M. M. Amaral[a], Cinthia Obladen de Almendra Freitas[b] and Flávio Bortolozzi[a]

[a] Centro Universitário de Maringá, Av. Guedner 1610, Maringá, Brazil
[b] Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição, 1155, Curitiba, Brazil

ABSTRACT

This paper describes an approach to writer identification based on graphometric features. These features are used by Forensic Document Examiners (FDE) which realize their analysis observing and extracting from the questioned documents a set of important individualizing features. Thus, in this work we present a baseline system composed of the following features: relative placement habits, relative relationship between individual word heights and relative slant. This set of features is submitted to Support Vector Machine (SVM) classifier as a writer identification method. The experimental results show accuracy rate equal to 80% considering all-against-all applying 100 writers. Finally, the obtained results are promising when compared to the literature.
.

Keywords: Handwriting recognition, forensic, document image processing.

1. INTRODUCTION

According to Morris[1], the forensic handwriting identification is part of criminology and its analysis provides a great number of elements related to personnel writing. The crime of forgery was established in the sixteenth century and the FDE have a hard task since time. In this context, computer-based methods and techniques have been proposed to provide support for this task.

Usually, the forensic handwriting identification is performed by experts using optical device and/or chemicals methods. The manual features extraction process can provide doubts about the writer identification[2]. In addition, different examiners can extract the same features from a particular document in a different way. Then, the use of semi-automatic systems can be useful and helpful to the experts when the problem is to identify the writer's identity.

In this context, different approaches have been presented in researches[3–8]. In these approaches, an important aspect is the feature set used to reveal the individual characteristics in the handwriting. Sreeraj and Idicula[9] present a classification which groups the features according to their granularity. Features which consider information from the entire document are classified as global, and features which consider information from a specific part of the document are classified as local. Based on the input method of writing, automatic writer identification system has been classified as online and offline[9].

Our work proposes an offline baseline system for writer identification, based only on graphometric features that are used by experts during their analyses. These features are extracted from different levels, such as document, line and word. Comparing with our previous research[10], this work includes the axial slant as part of the feature set, and the experiments were held applying a group of 100 different writers. These conditions have conducted to promising results.

This paper is organized as follows. Section 2 describes the proposed framework. Section 3 presents: the experimental results achieved and a discussion based on obtained results. Finally, Section 4 presents some considerations as well as points to future works.

2. FRAMEWORK FOR WRITING IDENTIFICATION USING GRAPHOMETRY FEATURES

Based on Sreeraj and Idicula[9], as mentioned before, approaches related to the feature extraction for writer identification can be divided into: global and local. Different approaches for offline writer identification have been presented in the literature. Many of them use features extracted from the document image, such as the texture approaches[11] or codebook approaches[5,8]. In our approach, as presented in researches[12-15], we applied only graphometric features, those applied by the FDEs.

Our baseline system is based on our previous work, as presented by Amaral et al.[10], adding as a new feature the axial slant. This framework is composed of the following steps: *Preprocessing* (thresholding, lines segmentation, first word of each line segmentation, contours extraction and document image implicit segmentation), *Feature Extraction* and *Classification*. To conduct our experiments we use 03 letters from 100 different writers (totalizing 300 different letters) from Brazilian Forensic Letter Database[17].

2.1. Feature Extraction

Based on graphometry, the feature set applied to writer identification process presented in this work is composed of: relative placement habits (f_1,f_3,f_4,f_5,f_6), relative relationship between individual words height (f_2 and f_7) and axial slant (f_8); as presented in Table 1.

Table 1. Feature set description

Group of Features	Description
f_1 (1)	Number of lines in forensic letter.
f_2 (2 to 21)	Proportion of black pixels: For the first 20 lines of the forensic letter computes the proportion of black pixels. The first word of each line is inserted in an implicit bounding box and later is computed the number of black pixels.
f_3 (22 to 41)	Right margin position: For the first 20 lines of the forensic letter is measured the distance between right margin and handwriting.
f_4 (42)	Lower left margin position: this distance is defined using the reference line and verifying the lower initial line position.
f_5 (43)	Upper margin position: this distance is defined by the first black pixel of the letter´s first word.
f_6 (44)	Bottom margin position: this distance is defined using the reference line and verifying the last line position.
f_7 (45 to 64)	Height of the first word: For the first 20 lines of the forensic letter is measured the height of the first word.
$f8$ (65 to 81)	Axial slant: directional histogram of the handwriting.

The result of the extraction process is a vector containing 81 primitives. This vector is applied to SVM classifier[18] in the training and testing stages. Figure 1 present an overview of the extraction process (f_1-,f_7) considering the image of a forensic letter. In this Figure it is possible to observe that lines and words of the segmented image are delimited and identified. Thus, information about the number of lines, margin positions and height/number of pixels of the words can be computed.

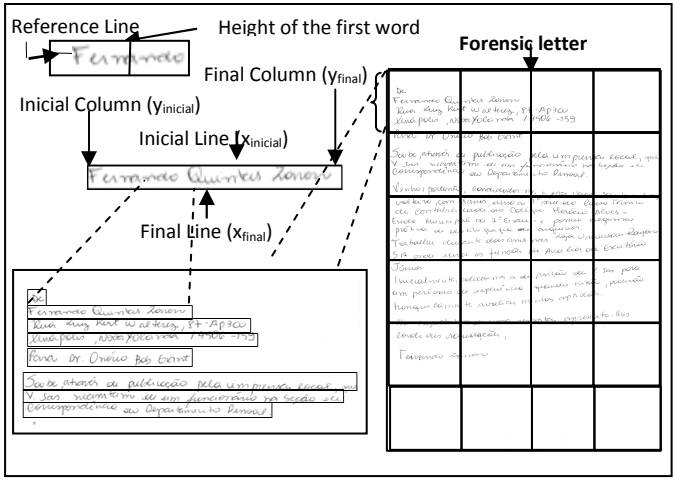


Figure 1. Overview of the features extraction – (f_1,f_2,f_3,f_4,f_5,f_6,f_7)

Individual handwriting characteristics are the particular character formed between movement and immobility during the act of writing, or aggregate of qualities, that distinguishes one person from others[1].

Taking this consideration, an important feature related to the handwriting individuality is the relative placement habits[1]. Writers may have a better use of the paper sheet, writing to your physical limit, but may also leave a blank space, usually regular in all lines. Different writers start and stop their writing at different locations. Then, locations, such as sentence indentation, shape of margins, use of space, starting and stopping points; are examples of the relative placement habits[1] (f_1,f_3,f_4,f_5,f_6).

Another important feature is related to the size of the first word of each handwriting line. In this work, the first word of each line was bounded by a box and its height and proportion of black pixels were computed (f_2, f_7).

Figure 2 present an overview of the axial slant (f_8) extraction process considering the image of a forensic letter. The axial slant is a graphometry feature used by the FDEs and has been extensively used in approaches to automatic writer identification[7]. This feature represents the general angle of the handwriting and has the best individual performance in the writer identification process proposed in this work, as demonstrated in Table 3.

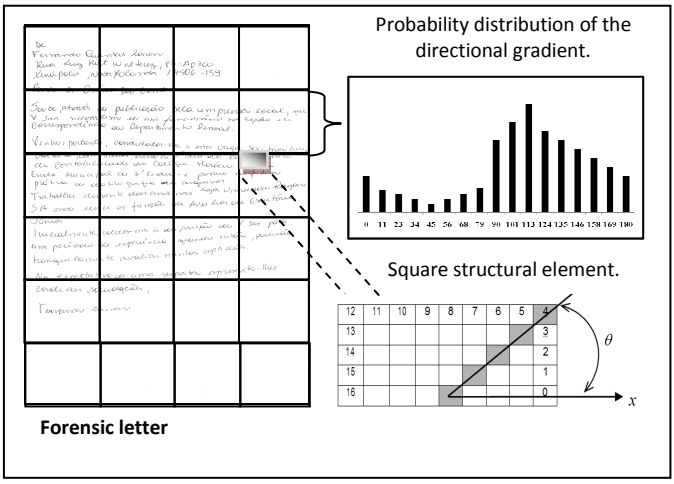


Figure 2. Overview of the feature extraction – (f_8)

In order to compute the axial slant feature, five segments are randomly selected from the segmented image (24 = 6 x 4). For each segment, its angle was computed[12]: all the directional gradients L (angles θ) are verified (Figure 2). This directional gradient vector is normalized by the probability distribution $P(\theta)$. The resulting histogram of each segment was added in the primitive vector submitted to the SVM classifier.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The classification method used in this work is SVM[18]. Besides, we applied the SMO (Sequential Minimal Optimization) algorithm[19]. SMO is a SVM implementation for training a support vector classifier using polynomial or RBF (Radial Basis Function) kernels. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default.

The feature vector is applied as input to the SVM. For the training stage this vector is used to compute the writer model, and for the testing stage the vector is used to writer identification. Two letters from each writer was used in the training stage as reference. At second stage (test stage), the framework compares a specific writer against the models obtained in the training stage, applying the third sample letter of each writer (all-against-all), so the best obtained result (winner-takes-all) is chosen by the SVM classifier.

Performing experimental results we used 100 writers (200 letters for training and 100 letters for testing, totalizing 300 different letters). The experimental results achieved accuracy rate equal to 80% demonstrating results comparable to those presented in the literature, as shown in Table 2.

Table 2 presents a comparison between our approach and other authors focused on graphometry for writer identification considering: features, classification methodology, number of writers used in the experiments and accuracy reached. An important result observed is the accuracy rate maintenance in our experiments with a greater number of writers (from 20 writers to 100 writers), due the addition of a new important feature in our baseline system: axial slant.

Table 3 demonstrates the identification rate distributed according to each group of features. This Table presents the accuracy rate of each feature, and some ensemble of features empirically defined. However, in a recent study, under development, a features selection process is been used to determine the best group of features.

It can be observed that the best ensemble is composed of $f_1\&f_6\&f_8$. Although the numbers of lines in each forensic letter (f_1) and bottom margin position (f_6) are not discriminatory features when applied as isolated feature. These features when combined to the axial slant *(f8)* allow the baseline system improving the final identification rate.

Table 2. Comparison among recent studies

Author	Number of writers	Feature	Classification Methodology	Accuracy (%)
Zois and Anastassopoulos[13]	50	Use of morphological operators to obtain the horizontal profile of the words.	Neural Network	95
Hertel and Bunke[14]	50	Continuity of the stroke, closed regions, upper and lower edges.	KNN	90
Schlapbach and Bunke[15]	50	Axial slant, height and slant of the text lines.	HMM	94.4
Pervouchine e Leedham[4]	165	Features extracted from the characters: “d”, “y” e “f” and the grapheme “th”.	DistAl Algorithm	58
Chen et al.[16]	60	Contour of adjacent segments.	SVM	54.9
Luna et al.[7]	30	Left margin and right margin positions, percentage space of the separation between lines, axial slant, space between words, proportion of words and the words slant.	ALVOT algorithm	92
Amaral et al.[10]	**20**	Number of the letter lines, proportion of black pixels, right margin position, lower left margin position, upper margin position, bottom margin position, height of the first word.	SVM	**80**
f_1,f_6,f_8	**100**	Number of the letter lines, lower left margin position and axial slant.	SVM	**80**

It is important to observe that the achieved accuracy rate using 20 writers obtained in Amaral et al.[10] (features f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7) decreases to

33% for 100 writers and achieves 60% when the axial slant (f_8) is added to the feature set.

Table 3. Writer identification performance (100 writers)

Features	**Accuracy** (%)	**Features**	**Accuracy** (%)
f_1	3	f_7	11
f_2	12	f_8	68
f_3	12	f_6 & f_8	77
f_4	5	f_1 & f_2 & f_3& f_4& f_5& f_6& f_7	33
f_5	5	f_1 & f_2 & f_3& f_4& f_5& f_6& f_7&f_8	60
f_6	5	**f_1 & f_6 & f_8**	**80**

This kind of experimentation is very important to confirm all the principles of handwriting and how the axial slant can be more than a simple writer habit but demonstrated that everyone's writing has a particular overall slant as explained by Morris[1]. It is important to detach that all studies presented in Table 2, which achieved best result than ours, used an inferior number of writers in their experiments, and in works as presented in this paper, the number of writers is fundamental to the reliability of the results.

4. CONCLUSION

In this paper we have discussed the efficiency of a graphometric feature set which can be applied to writer identification. Firstly, we have described the main features of graphometry and research works related to them. Thereafter, we presented the proposed method applied to the automatic feature set extraction. We have demonstrated, based on experimental results, using SVM classifier that these features achieved promising results for forensic handwriting analysis.

We observed that the set of features: number of lines in forensic letter (f_1), bottom margin position (f_6) and axial slant (f_8); was capable to perform the identification rate of 80% considering 100 different writers (all-against-all). As future work, new features will be studied and included in our baseline system trying to improve the results and some experiments based on different classifiers are been prepared.

REFERENCES

1. R. N. Morris. *Forensic Handwriting Identification – fundamental concepts and principles*. Academic Press, 2000.
2. G. Sheikholeslami, S. N. Srihari, V. Govindaraju. Computer aided graphology. in *Proceedings of the 5th International Workshop on Frontiers in Handwriting Recognition*, pp.457-460, 1996.
3. R. Plamondon, G. Lorrete. Automatic signature verification and writer identification: the state of the art. *Pattern Recognition*, vol. 37, n.2, pp. 107-131, 1989.

4. I. Siddiqi, N. Vincent. Combining global and local features for writer identification. in *Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition*, pp. 48-53, 2008.
5. Z. He, X. You, Y. Tang. Writer identification of Chinese handrwriting documents using hidden Markov tree model. *Pattern Recognition*, vol.41, pp.1295-1307, 2008.
6. B. Helli, E. Moghaddam. A text-independent Persian writer identification based on feature relation graph (FRG). *Pattern Recognition*, vol.43, pp.2199-2209, 2010.
7. E. C. H. Luna, E. M. F. Riveron, S. G. Calderon. A supervisoned algorihm with a new differentiated-weighting scheme for identifying the author of a handwritten text. *Pattern Recognition Letters*, vol.32, pp. 1139-1144, 2011.
8. L. Schomaker, K. Franke, M. Bulacu. Using codebooks of fragmented connected-component contours in forensic and historic writer identification. *Pattern Recognition Letters*, vol. 28, pp.719–727, 2007.
9. M. Sreejaj, S. M. Idicula. A survey on writer identification schemas. *International Journal of Computer Applications*, vol. 26, n. 2, pp.23-33, 2011.
10. A. M. M. M. Amaral, C. O. A. Freitas, F. Bortolozzi. The Graphometry applied to writer identification. in *Proceedings of the 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition*, Las Vegas, USA, vol.1, pp.10-16, 2012.
11. B. Helli, E. A. Moghaddam .A text-independent Persian writer identification based on feature relation graph (FRG). *Pattern Recognition*, vol. 43, p.2199-2209, 2010.
12. M. Bulacu, L. Shomaker, L. Vuupirjl. Writer Identification Using Edge-Based Directional Features. in *Proceedings of 7th Int. Conf.on Document Analysis and Recognition (ICDAR 2003)*, IEEE Press, 2003, pp.937-941.
13. E. Zois, V. Anastassopoulos. Morphological Waveform coding for writer identification. *Pattern Recognition*, vol. 33, n.3, pp. 385-398, 2000.
14. C. Hertel, H. Bunke. A set of novel features for writer identification. in *AVBPA'03 Proceedings of the 4th international conference on Audio- and video-based biometric person authentication*, pp. 679-687, 2003.
15. A. Schlapbach, H. Bunke. Off-line Handwriting Identification Using HMM Based Recognizers. in *Proceedings of the Pattern Recognition, 17th International Conference on ICPR'04* , vol.2, 2004.
16. J. Chen, D. Lopresti, E. Kavallieratou. The Impact of Ruling Lines on Writer Identification. in *12th International Conference on Frontiers in Handwriting Recognition*, 2010.
17. C. O. A. Freitas, L. S. Oliveira; F. Bortolozzi, R. Sabourin. Brazilian Forensic Letter Database. in *Proceedings of the 11th International Workshop on Frontiers on Handwriting Recognition*, 2008.
18. V. Vapnik. *Estimation of Dependences based on empirical data*. Nauka, Moscow, 1979. English translation: Springer Verlag, New York, 1982.
19. J.Platt, N. Cristanini, J. Shawe-Taylor. *Large margin DAGs for multiclass classification*. Advances in Neural Information Processing Systems. MIT Press. 2000.

Old document image denoising using bilateral filter

Mohamed Aymen Charrada[a] and Najoua Essoukri Ben Amara[b]

[a,b] Advanced System in Electrical Engineering Research unit (SAGE),
National Engineering School of Sousse (ENISo), University of Sousse, Tunisia
[a] mohamed_aymen_charrada@yahoo.fr
[b] najoua.benamara@eniso.rnu.tn

ABSTRACT

In this paper, we give an overview on old degraded image contrast enhancement and denoising approaches. We present also our contribution to this area, operating on the historical Tunisian heritage. The developed method is based on a novel bilateral filtering using a weighed k-means algorithm with a Gaussian kernel function for image denoising. It allows firstly the enhancement of the image contrast then it ensures the reducing of the background noise. The obtained results and the performed comparisons using our image database show many interesting perspectives.

Keywords: Old documents, denoising, contrast enhancement, k-means, bilateral filter

1. INTRODUCTION

Over a long period of human history, publication and transmission of information has been provided by the paper which represented the main support of knowledge transfer and communication from one era to another. However, this support has often suffered from an enormous fragility which caused a huge loss of the human heritage. This fact and following the scientific renaissance, community has proposed many innovative technological means allowing the conservation and the protection of the heritage saved on the paper. Indeed, the idea has been to digitize the patrimonial documents and to preserve them on the computer supports which provided new means for the retrieval, the recovery and the storage of the heritage information.

However, the majority of the archive documents dates from ancient time and suffers from many problems related mainly to the quality and the clarity of information and caused by the poor conditions of storage and conservation in the archives. In this sense, the digitalization operation must cohabit with specific treatments intended to restore the historical documents. Therefore, the selected digitalization mode extends to a processing chain which allows, firstly, to fill the

gaps previously mentioned and, secondly, to ensure the correction, the cleaning of the document and the improvement of its quality in order to guarantee better conditions of access and to provide the digital version most faithful to the original document.

In this article, we focus on the problems related to the preprocessing of degraded historical documents, our objective is to develop methods for denoising and contrast enhancement adapted to the level of difficulty present in the considered image database.

In the next section, we present a state of the main approaches of degraded historical documents denoising and contrast enhancement. Then we describe in Section 3 the proposed approach. We present, in section 4, the recorded experimental results and a comparison of our method with other denoising approaches.

2. PREVIEW ON THE OLD DOCUMENT IMAGE ENHANCEMENT

The study of ancient documents shows generally the existence of several degradations which can either be introduced by capture tools during the digitalization process such as the lighting variation, the inclinations, the curvatures, the blurred edges, the parasitic points ... or intrinsic to the documents and caused by the bad conditions of conservation and storage such as humidity, acidity, folds and tears, bleed-through... [1,2], which can taint the contents of these documents and affect their clarity and quality. A study was conducted on our image database in order to determine the occurrence percentages of the principal degradation classes. Table 1 presents the main obtained results.

Table 1. Principal degradation occurrence percentages on our database images

Degradations	Percentages
Low contrast	53%
Bleed-through	41%
Background noise	87%
Humidity	24%
Burrs & tears	28%
Acidity	19%

Among these degradations, the low contrast between the background and the text and the presence of the noise are two problems which appear frequently in the images of ancient documents and which requires the implementation of adapted preprocessing operations.

Indeed, several researches have led to the proposal of a set of preprocessing techniques in order to increase the contrast and remove the background noise

from the historical document images. Table 1 and Table 2 present a set of techniques that have been advocated in the literature to solve these two degradations.

Table 2. Selection of contrast enhancement methods suggested in literature

Reference	Approach description
[3]	Image histogram normalization
[4]	Spatial filtering technique and gray scale mathematical morphology
[5]	Image histogram Stretching and thresholding
[6]	Adaptive histogram equalization and bilinear interpolation
[7]	Image rescheduling

Table 3. Selection of denoising methods suggested in literature

Reference	Approach description
[5]	Wavelet decomposition and local thresholding (Donoho threshold)
[8]	Total Variation regularization and Non-Local Means filtering
[9]	Gradient-based method and Orientation-Isotropy Adaptive Filtering
[10]	Nonlinear blind source separation (BSS) algorithm
[11]	Hyperspectral imaging (HSI)
[12]	Nonsubsampled Contourlet Transform (NSCT)

We present in the third section our contribution to the enhancement of the degraded historical document images, which represents a color image denoising technique based essentially on the use of the bilateral filters.

3. PROPOSED APPROACH FOR ANCIENT DOCUMENT DENOISING

As part of our ancient document restoration project and in collaboration with the National Archives of Tunisia (NAT), we propose a new contrast enhancement and denoising approach for the historical documents based on the use of bilateral filters. This approach improves the image quality in order to proceed to the

subsequent phases of treatment with an aim of restoring these images. In the following, we begin with a presentation of the bilateral filter.

3.1 Bilateral filter

In fact, the bilateral filter, which is introduced by Tomasi and Manduchi in 1998, is an edge-preserving, a nonlinear diffusion and a noise reducing smoothing filter. The intensity value at each pixel in an image is replaced by a weighted average of intensity values from nearby pixels. This weight is based on a Gaussian distribution. Crucially, the weights depend not only on Euclidean distance but also on the radiometric differences (color intensity or Z distance). This preserves edges by systematically looping through each pixel and adjusting weights to the adjacent pixels accordingly [13]. Thus, if "p" is a pixel of the original image "I" and "I_R" is the restored image then the bilateral filter can be written as follows:

$$I_R(p) = \frac{1}{W_p} \sum_{x \in V(p)} G_{\sigma_s}(|x - p|) \cdot G_{\sigma_r}(|I(x) - I(p)|) \cdot I(x) \quad (1)$$

With:

$$W_p = \sum_{x \in V(p)} G_{\sigma_s}(|x - p|) \cdot G_{\sigma_r}(|I(x) - I(p)|) \quad (2)$$

$$G_\sigma(k) = exp\left(\frac{k^2}{2\sigma^2}\right)$$

And:

- $\frac{1}{W_p}$: Normalisation factor
- $V(p)$: Neighborhood of the pixel p
- G_{σ_s} : Space weight
- σ_s : Spatial extent of the kernel, size of the considered neighborhood
- G_{σ_r} : Range weight
- σ_r : Minimum amplitude of an edge
- $I(x) - I(p)$: Intensity difference for grey-level images and Color difference for color images

3.2 Proposed approach

The literature shows that several denoising techniques are based on the exploration of this filter [13]. Indeed, it offers a good opportunity for the reduction of the noise in the background keeping intact the details and the general structure of the image. To implement our proposed filter, we have replaced the original Euclidean distance, used by the bilateral filter, by a new kernel-induced distance function. According to the experiments, the standard bilateral filter cannot remove the noise in an effective way; this is mainly due to the fact that the noised pixel intensities are much higher or much lower than the intensities of their neighboring pixels so the Euclidean distance cannot provide a good estimator i.e. the initial estimator (I(x) in the equation (1)) used for the calculation of $I_R(p)$ is far from its real value, so the noised pixels remain almost

unchanged. To remedy this problem, we have replaced the Euclidean distance by a weighed k-means clustering function to define our new adaptive bilateral filter and to estimate the near-true values of I(x) which has allowed to calculate a number of weighted averaging of values among similar values. Figure 1 describes our contrast enhancement and denoising approach:

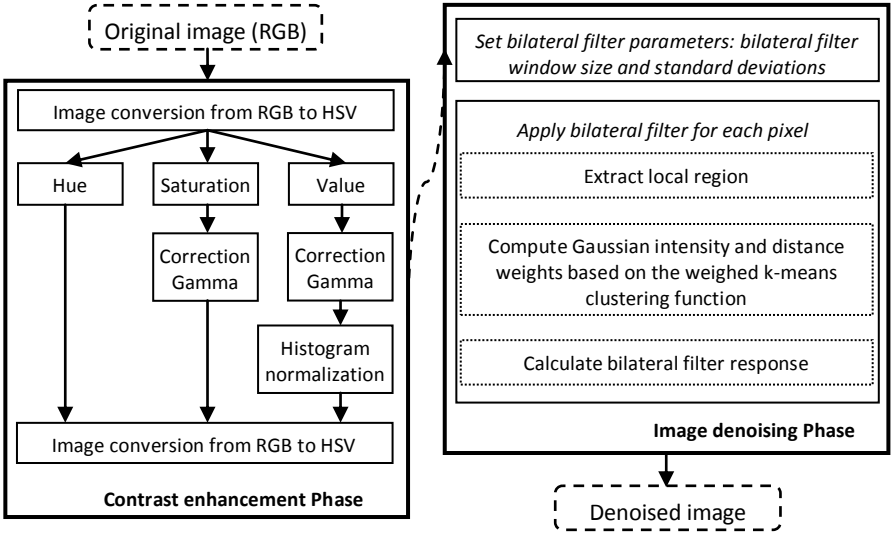


Figure 1: Flowchart of the proposed approach

In fact, after applying the contrast enhancement algorithm, our denoising approach is based on the application of the adapted bilateral filter for each pixel of the image as follows:

- Step 1: Extract the neighborhood of the corresponding pixel (local region);
- Step 2: Apply the k-means algorithm proposed in [14] on the pixels of the local region. This version of "k-means" allows to determine automatically the value of "k" based on the intra-cluster and inter-cluster distance measures;
- Step 3: Change the value of neighboring pixel "I(x)" in equations (1) and (2) by the value of its membership class center "I(c)" and calculate the filter response;
- Step 4: Repeat the second and the third steps for all the image pixels.

4. EXPERIMENTS AND RESULTS

We used an image database consisting of manuscripts in Arabic and French languages dating from 1547 to 1956, coming from the NAT. This database consists of about 10000 historical document images characterized by a great structure variability and content richness. The considered documents are with

variable size, scanned in color mode at a resolution of 300 dpi using the “JPEG” format. This database is described in [15].

In order to evaluate our approach, we have compared, qualitatively and quantitatively, its performance with two state-of-the art filtering algorithms: an Orientation-Isotropy Adaptive Filtering approach [9] and a Non-Local Means filtering approach [8]. All experiments were performed using MATLAB 7.4.0. Figure 2 illustrates the results of this comparison.

Figure. 2 : Comparison of denoising results obtained by the three approaches: (1) Original images, (2) Our approach, (3) Approach presented in [9], (4) Approach presented in [8]

To measure the quality of filters, Table 4 shows the average values of the Peak Signal-to-Noise Ratio (PSNR) and the Root Mean Squared Error (RMSE)

recorded by the three denoising approaches. The best technique allows to have the lower RMSE value, and the higher PSNR value.

Table 4. Comparison of the Recorded results

Approaches	**Results**					
	Image 1		***Image 2***		***Image 3***	
	PSNR	**RMSE**	**PSNR**	**RMSE**	**PSNR**	**RMSE**
[9]	29.18	13.92	34.63	10.07	34.06	12.73
[8]	33.59	14.51	**37.26**	7.32	36.56	10.19
Our method	**34.40**	**7.56**	36.62	**5.82**	**37.82**	**5.08**

Figure 2 and Table 4 show that our approach provides better denoising results compared to the other tested methods. In fact, our method preserves edges, corners and high frequency characteristics of the filtered image as it provides good results in the presence of texture; it does not modify, in the most of cases, the structure and the details of the image as it reduces greatly the loss of the useful information. Also, our method is less sensitive to the variation of luminance, the low contrast and the presence of other types of degradations compared to the other considered methods.

5. CONCLUSIONS AND PROSPECTS

In this paper, we have presented a new method for old color image denoising. This method allows firstly the increase of the contrast using an adaptive adjustment and standardization process for the image components (Hue, Saturation and Value). In the second place, it reduces the background noise in the processed images by exploiting the bilateral filter with a Gaussian distribution. In fact, this method allows to denoise images preserving their main structures and their details. The comparison of our approach with other denoising techniques, proposed in literature, shows that the obtained results are encouraging. However, tests are underway to optimize further the proposed denoising procedure.

ACKNOWLEDGMENTS

Special thanks to the Tunisian National Archives (NAT) to have given us access to their large image database of Tunisian historical documents.

REFERENCES

1. A. Ghardallou, *Apport des Outils de Traitement de Signal en Numérisation des Documents Anciens*, Master Degree, Faculty of Sciences of Monastir, pp.14-18, June 2006.

2. F. Drira, *Contribution à la Restauration des Images de Documents Anciens*, PhD degree, National Institute of Applied Sciences of Lyon, pp.16-38, 2007.
3. W. Boussellaa, A. Zahour and A. Alimi, "A Methodology for the Separation of Foreground/Background in Arabic Historical Manuscripts using Hybrid Methods", *Journal of Universal Computer Science*, **Vol. 14, No. 2**, 2008.
4. B. Gangamma, K. Srikanta Murthy and A. Vikas Singh, "Restoration of Degraded Historical Document Image", *Journal of Emerging Trends in Computing and Information Sciences*, **Vol. 3, No. 5**, May 2012.
5. A. Kricha, *Contribution au prétraitement et à la caractérisation des documents anciens : Application à la segmentation*, PhD degree, March 2013.
6. E. Kavallieratou, "A Binarization Algorithm specialized on Document Images and Photos", *in International Conference on Document Analysis and Recognition (ICDAR)*, pp 463-467, 2005.
7. T. Obafemi-Ajayi, G. Agam, and O. Frieder, "Ensemble LUT classification for degraded document enhancement", *in Document Recognition and Retrieval XV*, Proc. SPIE **6815**, 2008.
8. L. Likforman-Sulem, J. Darbon and E.H.B. Smith, "Enhancement of historical printed document images by combining Total Variation regularization and Non-local Means filtering", *Image and Vision Computing*, **Vol. 29, No. 5**, pp. 351-363, 2011.
9. H. Wang, Y. Chen, T. Fang, J. Tyan and N. Ahuja, "Gradient Adaptive Image Restoration and Enhancement", *in ICIP*, pp. 2893-2896, 2006.
10. F. Merrikh-Bayat, M. Babaie-Zadeh and C. Jutten, "A Nonlinear Blind Source Separation Solution For Removing The Show-Through Effect In The Scanned Documents", *in Proc. 16th European Signal Processing Conference (EUSIPCO-2008)*, Lausanne, Switzerland, 2008.
11. S.J. Kim, F. Deng and M. S. Brown, "Visual Enhancement of Old Documents with Hyperspectral Imaging", *in Pattern Recognition*, **Vol. 44, No. 7**, July, 2011.
12. Z. Hua, Y. Li and J. Li, "Image Nonlinear Enhancement Algorithm based on Nonsubsampled Contourlet Transform", *In Journal of Digital Content Technology and its Applications*. **Vol. 5, No 7**, July 2011.
13. G.Vijaya and V.Vasudevan, "Bilateral Filtering using Modified Fuzzy Clustering for Image Denoising", *In Journal on Computer Science and Engineering (IJCSE)*, **Vol. 3, No. 1**, pp. 45-49, 2011.
14. S. Ray and R.H. Turi, "Determination of Number of Clusters in K-means Clustering and Application in Color Image Segmentation", *In Proc. of the 4th Intern. Conf. on Advances in Pattern Recognition and Digital Techniques*, India, 1999.
15. M.A. Charrada, N. Essoukri Ben Amara, "Development of a database with ground truth for old documents analysis", *In Proc. of the International Multi-Conference on Systems, Signals and Devices (SSD'2013)*, Hammamet, Tunisia, 18 - 21, March 2013.

Efficient Binarization-Free Text Line Segmentation for Historical Documents

Angelika Garz
University of Fribourg, DIVA
1700 Fribourg, Switzerland
angelika.garz@unifr.ch

***Abstract*—Segmenting page images into text lines is a crucial pre-processing step for automated reading of historical documents, and is still an open research topic due to challenges such as heterogeneous and noisy background, ink bleed-through, artifacts due to aging, stains, and touching text lines. We present a binarization-free line segmentation method robust to noise, which copes with overlapping and touching text lines. It is a bottom-up approach based on clustering interest points representing parts of characters to words. Then touching components such as ascenders and descenders are split using seam carving. Text lines are generated by concatenating neighboring word clusters. An experimental evaluation on images of the Saint Gall database shows promising results for real-world applications in terms of both accuracy and efficiency.**

I. Introduction

Various projects for digitizing historical documents have been realized in the last decades, resulting in large digital databases. In order to make these searchable and useful to scholars, contents have to be indexed and transcribed automatically. Handwriting recognition in historical documents, however, requires text line segmentation algorithms to be robust with respect to background artifacts such as clutter, stains and noise, as well as artifacts due to aging, and touching or interfering lines [2]. Handwritten documents do not have strict layout rules and thus line segmentation methods need to be invariant to layout inconsistencies, irregularities in script and writing style, skew, and fluctuating text lines [2]. Furthermore, robustness to low contrast and rippled pages is required [3].

While a detailed survey about text line segmentation with respect to historical documents is done by Likforman et al. [2], general research directions and recent approaches are summarized here. An established method for documents with constrained layouts are Projection Profiles (Pp) [3–7] for both binary and gray-scale images. Various authors [4, 5] adapted the global Pp such that skewed text blocks, converging or merging text lines are segmented correctly. Further methods for binary images include smearing [8, 9] and Hough transform [10, 11]. Seam carving [12] known from image retargeting, is applied by several recent approaches [13–15] Nikolaou and Gatos [15] use so-called local minima tracers which follow the line spacing in order to shred the document page into lines. Indermühle et al. [13] use Dynamic Programming (Dp) in order to find a path with the minimum cost between two lines in historical manuscripts. Asi et al. [14] apply their approach directly on gray-scale images, where a distance transform is computed from a Gaussian-blurred image, and the separating seams are established using Dp. Li et al. [16] propose a method based on density estimation and the level set method.

Of the methods published, the majority relies on binarization. The issue whether or not to use binarization has been discussed extensively in the literature [17, 18]. Valuable information encoded in gray-scale images is discarded in binary images. Furthermore, binarization of historical documents is prone to errors due to heterogeneous background, artifacts, and noise owing to materials used and aging processes. The information loss and errors introduced impair all further processing steps. Using seam carving on entire document images is computational intensive and requires knowledge about the number of text lines and their locations.

We present an efficient binarization-free method for line segmentation applicable to historical manuscripts. Our method transforms the page image into the domain of interest points (Ips); Ips represent parts of characters. The presented method follows a bottom-up concept; it groups Ips describing parts of characters into text lines. Touching components such as ascenders and descenders are locally split by means of seam carving. Prior layout analysis is not compulsively necessary and the approach is independent of any layout model. A previous detection of numbers and locations of text lines as done in [14] is not necessary. The approach is evaluated on pages of the *Saint Gall database*[1], which is part of IAM HistDB [19]. It consists of 60 pages of a Latin manuscript originating from the 9^{th} century written in Carolingian script by a single writer with ink on parchment. Two sample pages are illustrated in Figure 1.

The remainder of this paper is structured as follows. First, the proposed method is explained; experimental results are depicted in Section III, followed by conclusions drawn in Section IV.

II. Methodology

First Ips are extracted from gray-scale images by means of Difference of Gaussian (DoG), the Ip detector used for Scale-Invariant Feature Transform (Sift) [20] . Note that after extraction of Ips, line segmentation can be realized very

This paper is a summary of work [1] previously published at DAS2012.

[1]Available at `http://www.iam.unibe.ch/fki/databases`

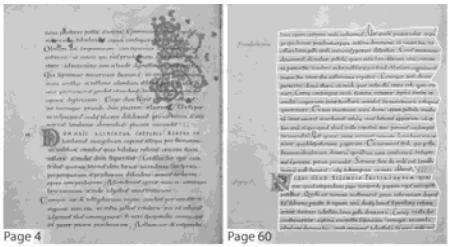


Figure 1. Samples of the *Saint Gall database*. Page 60 is overlaid with contours illustrating the ground truth. Pages have regular layouts with relatively wide text line spacing; contain colored initial letters and annotations. However, ascenders and descenders connect adjacent lines. Stains, holes, and ink bleed-through pose additional challenges.

efficiently in the sparse IP domain. IPs describe structures dissimilar to their adjacent neighborhood, e.g. in terms of intensity or color. An IP has a defined location in the image and a definite spatial extend. Layout analysis [21] prior to the segmentation of text lines is not a compulsive requirement due the very nature of the IPs extracted by means of DoG, which allows the selection of thresholds in both, the scale and the sensitivity in terms of pixel intensity. Thus, IPs are mostly detected on and between characters. In order to identify word clusters, spatial clustering robust to noise is applied to the IPs extracted. Adjoined words of consecutive text lines connected by touching ascenders and descenders are identified and separated using DP. Finally, text lines are generated connecting adjacent word clusters in the direction of the mean orientation of the text in the page. In the following, the successive steps of the proposed method are explained.

A. Identification of Word Clusters

Word clusters are identified in high-density regions, since IPs are mainly detected on and between characters and only few IPs are generated for background areas. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [22] is applied to the IPs' coordinates in order to group adjoining characters into word clusters. Unlike other clustering algorithms, such as k-means, the number of clusters is determined by neighborhood constraints and does not need to be defined. In order to identify a cluster, the density in the neighborhood has to exceed a certain threshold [22], i.e. a minimum number of neighboring data points has to be given along with a search radius. Thus, background noise is not included in word clusters. We manually estimate the search radius by defining it as half the x-height[2] (see Section III).

Then, for each word cluster, a minimum area rectangle is calculated such that one edge of the rectangle is aligned with an edge of the convex hull (see Figure 2 a). The text orientation of a page is calculated as the mean of the

[2]Height of a lower-case letter without ascenders and descenders.

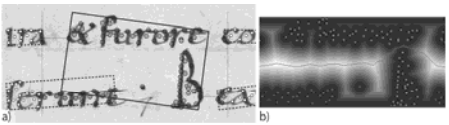


Figure 2. a) Image patch overlaid with markers indicating of IPs, and minimum area rectangles of the word clusters. Green markers represent IPs belonging to the word cluster to be split, orange markers indicate IPs of adjacent word clusters selected to provide context. Dotted orange lines illustrate the distance within which IPs of adjacent word clusters are regarded. Further IPs not taken into account are presented as white dots.
b) Orientation-normalized distance transform overlaid with the IPs, the boundary condition (red) and the path found (blue).

orientations of the word clusters, which is estimated by the median of the first principle component direction of the spatial distribution of IPs. We orientation-normalize the word clusters and automatically determine the median word height used for further steps, where the word height is defined as x-height.

B. Identification and Separation of Touching Components

Word clusters spreading over consecutive lines owing to touching components such as ascenders and descenders are separated by seam carving. They are identified based on a height threshold, which is mainly dependent on the automatically determined median word height, but needs some fine-tuning for a given manuscript in practice (see Section III). IPs of adjacent word clusters within a distance of the median word height are additionally included in order to embed the word cluster into a line context (see Figure 2 a, dotted orange lines); and a boundary condition is introduced which prevents the path from propagating into a local maximum at the border of the word cluster, which potentially happens if there is no adjacent word cluster in one of the text lines.

The energy function is calculated as distance transform of the IPs' coordinates and the boundary condition using Euclidean distance resulting in an orientation-normalized energy map (see Figure 2 b). DP is employed in order to find the path of maximum energy (farthest distance to the IPs and the boundary condition). Finally, the word cluster is split according to the calculated path.

C. Generation of Text Lines

The last step is generating text lines by concatenating word clusters. Each word cluster is connected to its nearest neighbors to the left and right in the mean text orientation by drawing a fictive line from the center of mass of a word cluster. A word cluster is selected if at least one corner of its rectangle is on the opposite site of the line, which allows for a fast check. Then, chains of connected word clusters are built and chains with the maximum number of word clusters are identified as text lines. Figure 3 illustrates an example of line generation.

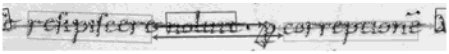


Figure 3. Concatenating word clusters. IPs are indicated by white dots, word clusters are surrounded by rectangles. The fictive lines originating in the centers of mass of each word cluster are indicated by arrows in the respective color. The blue word cluster p is below the mean line position, however, it is hit by its neighboring clusters (violet, green), and hits the orange and the green cluster to its left and right respectively.

Handwritten documents contain curved and fluctuating lines; however, we assume the orientation not to change abruptly from one word to the next. Thus, the probability of the adjacent word cluster being hit by locally applying the prevailing text orientation as search direction is high.

As a last step, the IPs are spatially weighted with a two-dimensional Gaussian distribution with a standard deviation according to their scales (see [21]) in order to generate contours of the text lines generating a probability map for each text line. These probability maps are voted against each other with the higher probability determining the text line a pixel belongs to, resulting in text line regions (see Figure 4).

III. Experimental Evaluation

The evaluation is performed on all pages of the *Saint Gall database*. First, a randomly selected page was used to fine-tune system parameters. Most importantly, the threshold on the scale and sensitivity of the DoG IPs needs to be optimized (see Section II). Other parameters include the neighborhood radius for DBSCAN (see Section II-A) and the threshold used for separating touching components (see Section II-B). Although reasonable defaults can be set for these parameters with respect to the automatically determined median word height, some fine-tuning improves the results in practice.

Figure 4 gives a sample result, where detected text lines are represented by their contours, which are randomly colored for easier distinction. Figure 5 shows an image patch of the manuscript with bleed through from the other side of the page. Nevertheless, the word clusters are found in presence of noise and lines are segmented correctly.

The quantitative performance evaluation is based on the measures *pixel-level hit rate* and *line accuracy* introduced by Li et al. [16]. A binarized version of the database is used for the evaluation procedure on pixel-level. Although the use of binarization ground truth is ambiguous and contains some background noise, the performance measure provides reasonable results which were verified by visual inspection. The *pixel-level hit rate* measures the performance on pixel-level, i.e. the amount of ground truth pixels that are retrieved with optimal assignment. The best assignment between a line detected by the system and a line in the ground truth is found based on corresponding pixels by means of the Hungarian algorithm [16]. Segmentation errors – such as splitting, merging, and missing lines – are penalized. The *line accuracy measure* evaluates the performance on text line level.

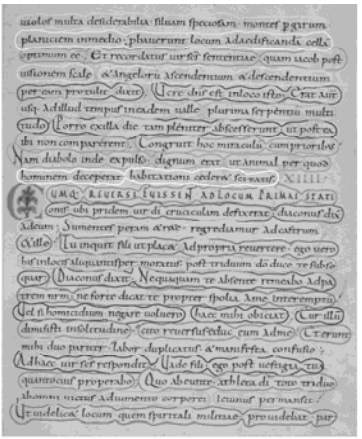


Figure 4. Page 21 overlaid with the contours of detected text lines.

Figure 5. Line segmentation in presence of background clutter.

A line in the ground truth is claimed to be detected, if more than $90\,\%$ of the pixels are shared with the corresponding detected line – with respect to both of them. Missing parts of text lines, noise or additionally added parts of other text lines are penalized [16].

We achieve a hit rate of 0.9865, which expresses the amount of ground truth pixels that are retrieved with an optimal assignment; and a line accuracy of 0.9797 on the *Saint Gall database*, which has a total of 1,431 text lines.

A. Typical Failures and Causes

Part of the error can be traced back to imperfect separation of word clusters containing touching components, since the separating path might cut an ascender or descender if the distance between two adjacent IPs on the component is larger than this between IPs of the two components (see Figure 6 a).

Since prior layout analysis is not done, IPs are found in initials and on the border of holes if their contrast to the background is high enough, which results in noise in the detected lines (see Figure 6 b,c).

In one case, a word cluster was not split due to the large amount of structure between the consecutive lines, resulting in a merge of consecutive text lines (see Figure 6 d). In another case, a word cluster spreading over two lines is not split since its height is below the threshold to be considered for a split (see Figure 6 e).

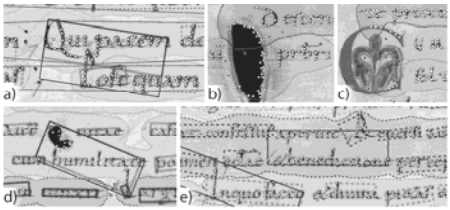


Figure 6. Image patches overlaid with markers representing IPs, minimum area rectangles color-coded according to the line they belong to, and colored contours indicating the text lines.
a) A word cluster (indicated by a black rectangle partly overlaid by the green one) is split imperfectly into two word clusters (orange and green).
b,c) IPs found on borders of holes and in initials.
d) Two merging lines due to a word cluster without a local minimum in the vertical distribution between two lines (letter *I* of the green line is split).
e) Two merging lines due to a word cluster having a vertical extend smaller than the threshold to split (see Section II-B).

B. Efficiency

Once having transferred the document image into the IP domain, methods are either applied on approximately 9,800 IPs (dependent on the actual image) or the minimum area rectangles of the word clusters instead of 16.6 million pixels, resulting in efficient processing.

Generating the energy function for seam carving is expensive due to dependencies of the pixels on each other. First identifying the word clusters which need to be split, we efficiently apply the seam carving approach only in a local area, thus further improving the efficiency of the method.

IV. Conclusion

We presented a binarization-free method capable of segmenting text lines in historical handwritten documents, which relies on IPs representing letters. Once IPs are extracted, they are applied to the sparse IP domain and thus, can be efficiently implemented. First, word clusters are identified in high-density regions using spatial clustering. Then word clusters containing touching or overlapping components are identified and separated by means of DP. Finally, adjacent word clusters in the direction of the mean text orientation are joined to lines.

Based on local information only, this approach is robust to shape, translation, and rotation of text blocks, and slight curvature of baselines. The method is not designed for a specific script and thus, can be applied to other scripts with the constraint that a set of system parameters needs to be fine-tuned for a sample page. The reported line segmentation accuracy for the Latin manuscripts of the *Saint Gall database* is promising for real-world handwriting recognition applications.

Since the publication of the original paper [1], the presented approach has been largely modified and further developed in order to cope with less contrast, and stronger curvature of text lines, and has been tested on synthetic data in order to evaluate its robustness. An according paper is currently under review.

References

[1] A. Garz, A. Fischer, R. Sablatnig, and H. Bunke, "Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering," in *DAS*, 2012, pp. 95–99.

[2] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text Line Segmentation of Historical Documents: A Survey," *IJDAR*, vol. 9, no. 2, pp. 123–138, 2007.

[3] M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant, "Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen," in *ICDAR*, vol. 1, 2007, pp. 357–361.

[4] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic Hand-Written Text-Line Extraction," in *ICDAR*, 2001, pp. 281–285.

[5] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein, "Line Segmentation for Degraded Handwritten Historical Documents," in *ICDAR*, 2009, pp. 1161–1165.

[6] A. Antonacopoulos and D. Karatzas, "Document Image Analysis for World War II Personal Records," in *DIAL*, 2004, pp. 336–341.

[7] H. Adiguzel, E. Sahin, and P. Duygulu, "A Hybrid for Line Segmentation in Handwritten Documents," in *ICFHR*, 2012, pp. 503–508.

[8] Z. Shi and V. Govindaraju, "Line Separation for Complex Document Images Using Fuzzy Runlength," in *DIAL*, 2004, pp. 306–312.

[9] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos, and N. Papamarkos, "Segmentation of Historical Machine-Printed Documents using Adaptive Run Length Smoothing and Skeleton Segmentation Paths," *Image a. Vision Computing*, vol. 28, no. 4, pp. 590–604, 2010.

[10] L. Likforman-Sulem, A. Hanimyan, and C. Faure, "A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents," in *ICDAR*, vol. 2, 1995, pp. 774–777.

[11] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text Line Detection in Handwritten Documents," *Pattern Recognition*, vol. 41, no. 12, pp. 3758–3772, 2008.

[12] S. Avidan and A. Shamir, "Seam Carving for Content-Aware Image Resizing," in *ACM Trans. on Graphics*, vol. 26, no. 3, 2007, p. 10.

[13] E. Indermühle, M. Liwicki, and H. Bunke, "Combining Alignment Results for Historical Handwritten Document Analysis," in *ICDAR*, 2009, pp. 1186–1190.

[14] A. Asi, R. Saabni, and J. El-Sana, "Text Line Segmentation for Gray Scale Historical Document Images," in *HIP Wksh.*, 2011, pp. 120–125.

[15] A. Nicolaou and B. Gatos, "Handwritten Text Line Segmentation by Shredding Text into its Lines," in *ICDAR*, 2009, pp. 626–630.

[16] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "Script-Independent Text Line Segmentation in Freestyle Handwritten Documents." *PAMI*, vol. 30, no. 8, pp. 1313–1329, 2008.

[17] D. Lopresti and G. Nagy, "When is a Problem Solved?" in *ICDAR*, 2011, pp. 32–36.

[18] G. Nagy, "Preprocessing Document Images by Resampling is Error Prone and Unnecessary," in *IS&T/SPIE Electronic Imaging 8658*, 2013.

[19] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription Alignment of Latin Manuscripts Using Hidden Markov Models," in *HIP Wksh.*, 2011, pp. 29–36.

[20] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[21] A. Garz, R. Sablatnig, and M. Diem, "Layout Analysis for Historical Manuscripts Using SIFT Features," in *ICDAR*, 2011, pp. 508–512.

[22] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *KDD*, 1996, pp. 226–231.

Detecting Main Body Size in Historical Document Images

Diamantatos Paraskevas Ergina Kavallieratou

dept. Information and Comunications Systems Engineering

University of the Aegean

Samos, Greece

oxinrain@gmail.com , kavallieratou@aegean.gr

Abstract

In this paper, we present a technique, appropriate to detect the text main body size in a historical document image. The proposed technique measures directly the main body size, without requiring image segmentation or binarization, although the results are more precise in pre processed binarized images. The proposed technique has the ability to correctly measure the main body size even if the document is slightly skewed.

Keywords—document image processing; word main body estimation; baseline detection; historical document images

1. Introduction

Main body or core region size is a characteristic that is used quite often in most document image processing systems. By this term, it is considered the central part of the text, excluding ascenders and descenders (Fig.1). Most of the times, it is referred to words.

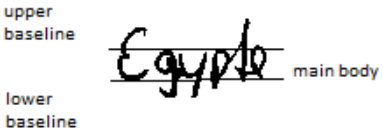


Fig.1: Word main body and baselines.

Main body is a characteristic used in a large number of systems that use image processing for a variety of tasks in document images. It has been used in systems for OCR [1-2], segmentation [3-4], slant removal [5], dewarping [6-7], word matching [8], indexing [9], normalization [10], word spotting [11], etc. All the above systems utilize the main body information, and use it as a threshold or character size information, as it is directly related to the size of the characters, the document image resolution and the text orientation.

Main body can also be utilized in order to get a rough estimation of character width. Especially in [14], they mention: *By mean width of character, we consider the width of characters such as a, b, c, d etc., excluding the characters i,l,j,m,w that are either too narrow (i,l), or too wide (m,w). Although the character width differs between characters and writers, a rough estimation of the mean width could be made by accepting that excluding the ascenders and descenders the characters with mean width (as defined above), present width equal to their height.*

Considering all the above, we see that main body is absolutely crucial in document image processing systems. Thus, many techniques have been developed for calculating main body.

In this paper, a simple technique is presented, low in computational cost. Moreover, it does not require the text baselines localization. Instead computes the main body directly.

In the next section a short description of the previous work is given, while the proposed technique is presented in section 3. In section 4 we present a way of finding an optimal threshold, on chapter 5 we describe some of the results, while in section 6, we conclude.

2. Previous work

Although there is no paper specific for main body estimation, in document image processing literature we can find a lot of suggested techniques for main body estimation.

Lee[5] and Adamek [12] are using pixel level processing. Lee measures distances between pixels, vertical transitions and try to find baselines from which we can get main body, while Adamek using pixel density to measure main body size.

Cheng [4] and Cote et al. [2] both are using histograms to measure main body size. Cote et al also calculates the entropy associated with the histograms. In more complex techniques Marti [1], Gatos et al. [8] and Sharma [9] use linear regression to estimate the upper and lower baselines and finally Papavassiliou et al. [15] formulate an HMM for the text and gap stripes within the document image.

While all the above techniques calculate main body size, they require word or line segmentation even if the required task of segmentation isn't desired. The technique we present in the next section doesn't require line segmentation.

3. The proposed technique

Our technique, shown briefly in fig.2 and analyzed in this chapter, estimates the average main body of words in a scanned document. Although it has some similarities with [6], it is not that complex, it does not require line segmentation nor image binarization. The technique is applied to grey level images, although the experimental results prove that if the image is binarized and cleaned from extra noise and then converted to grey level, the results are improved.

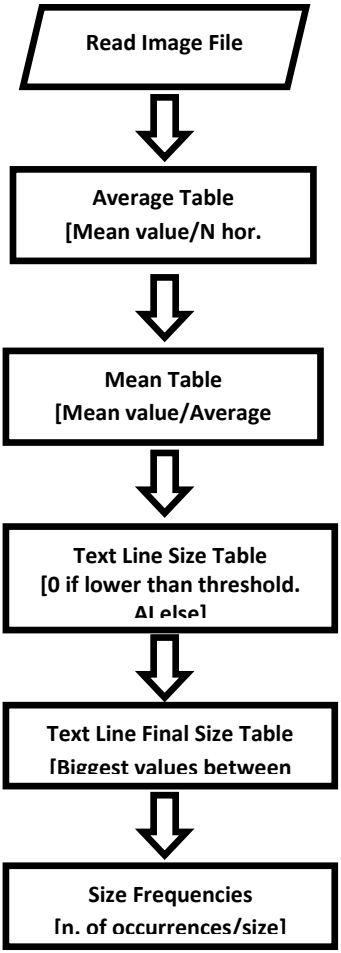


Fig.2: The proposed methodology.

First, the average pixel value for every consecutive N pixels of each pixel line is calculated. N can take any value, less or equal to the image width. This information is stored in the Average Table. In order to count the average main body for each text line, after calculating the average pixel values, we introduce the Mean table. This table can be skipped if the text lines have a big declination,

or if we are interested in getting the average main body of words. The Mean table stores a mean value per each average table line.

Next, a way to distinguish text lines is required. Empty lines are considered. Thus, we introduce a threshold. This threshold has to be determined for each image that hasn't been previously processed, and includes pixel values between 0 and 255. However, for the majority of the systems that require word main body, there is a binarization threshold. After several experiments, 230 was selected (explanation given below) as the threshold that can give very accurate results for the majority of document images. After applying this threshold to the mean, or to the average table, depending on the application, the Text Line Size Table is resulted. This table presents zeros (0) on the lines that are lower than the threshold, and an auto increment starts from 1 in the following lines, having a value bigger than the threshold.

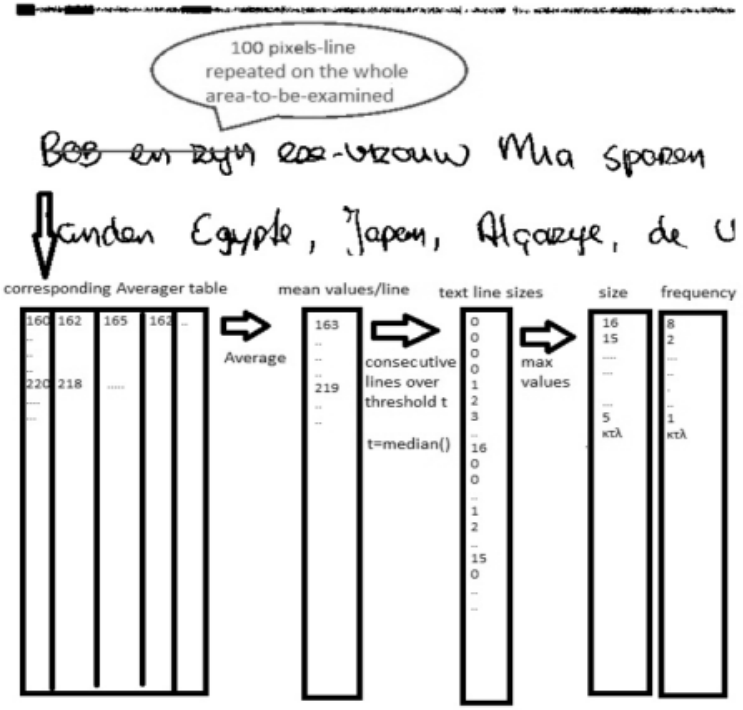


Fig.3 Schematic presentation of the technique through example

Finally, all the maximum values between zeros are considered and stored in the Text Line Final Size Table. Next, all the size occurrences are counted and stored in the size Frequencies table, along with their occurrence frequency. From this table the information about the various main body sizes is extracted, as well as, which one is the most common. An example of our technique is shown in fig.3. The ruled grey line symbolizes the consecutive N pixel Mean value . The first N

pixels mean value is stored in the first column (160), the next N pixels mean value is stored in second column (162) etc. Then the entire row (line) mean value is calculated and stored in the Mean value per line table (163). Then we calculate main body sizes along with their frequencies.

4. Finding the optimal threshold

One of the key points of this technique is selecting an optimal Threshold . While the experimental results proved that the same threshold (230) can be used for document images on the same data set that are first binarized and then converted to grey level , this does not apply to all document images. Furthermore if the document image hasn't been processed first , finding an optimal threshold isn't an easy task.

For the task of finding the optimal threshold we used several iterations of our technique using several different thresholds. In our experiments, it was set an auto increment threshold that on every iteration of our algorithm its value was increased by one. The threshold started from 100 and finished on 255. Also a max value for the returned main body size at 50 pixels and a lower value of 6 pixels was set (for 300 dpi). After several iterations we got 155 main body sizes and by calculating the size occurrences we got the result shown in fig.4.

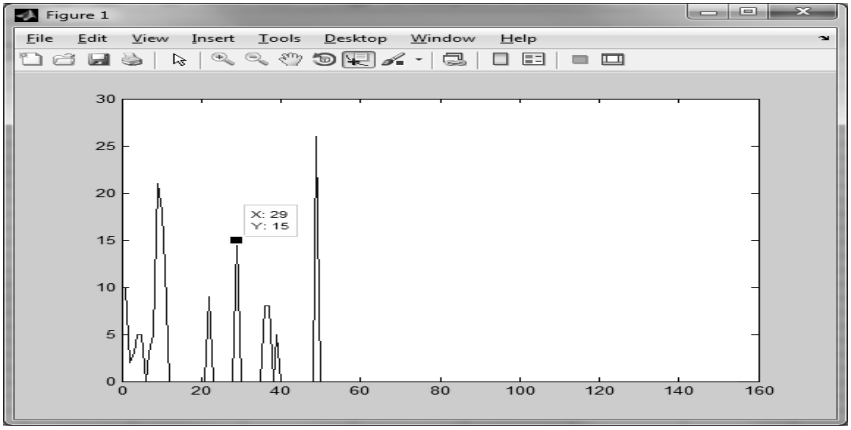


Fig.4 size occurrences while incrementing by 1 the threshold value on each iteration (X-axis: Main body size , Y-axis: occurrences)

The main body size should not be very close to 0 or 50, so it lies in the middle. The most common size between those values is 29 which is the same for threshold 230. This iteration process revealed also the size of each line, meaning the main body along with ascenders and descenders to be near 40. But while this solution can find the ideal threshold , it takes a lot more time.

While our technique needs a threshold estimation, it does not require image binarization or segmentation, and it can provide additional information for the detected main body sizes. Furthermore , in our experiments with TrigraphSlant db [16], after converting the images on grey level, the threshold value remained

the same for all the document images we tested. In fig.5 the detected main body sizes are shown along with their occurrences, while in fig.6 we can get additional information for every N pixels main body size occurrences. Finally on Fig.7 a sample output of our technique in matlab is presented for an image form TrigraphSlant data set with a vertical skew of 5 degrees made on paint.

5. Experimental Results

While the evaluation of a technique that estimates main body size is not an absolute task, especially for hand written documents, there are appropriate data sets like TrigraphSlant db [16] that contains images of handwriting, produced under conditions of natural and forced slant. This data set includes 190 images from 47 persons. For the purpose of the evaluation of our technique we used 30 images of natural writing by different writers, after estimating the main body size by human user. Our technique had 2.17 average error deviation in pixels.

6. Conclusion

The proposed technique has a very low average error deviation and has low computational cost, due to simple calculations. Although the given results seem good the evaluation technique we used doesn't give objective results. Future plans include a better evaluation technique.

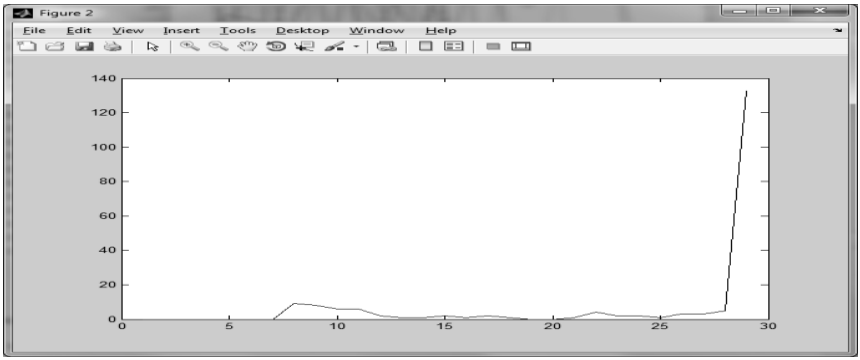


Fig.5 Plot with main body size occurrences(X-axis: Main body size , Y-axis: occurrences)

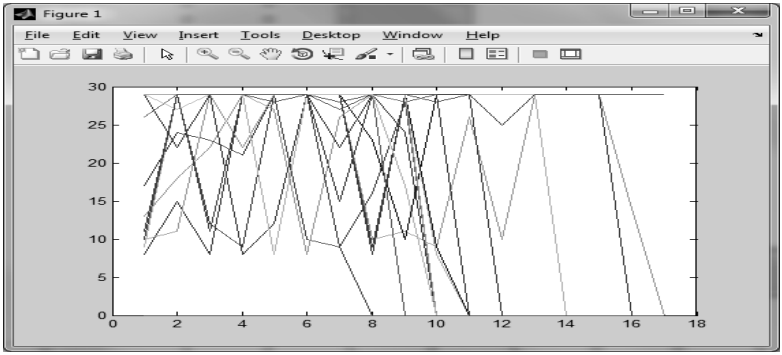


Fig.6 main body size occurrences for every N pixels (N= 100 for this example, X-axis: occurrences, Y-axis: Main body size)

```
>> megethos_leksis('D004skewV5.png',230)
Warning: Image is too big to fit on screen; displaying at 25%
> In imuitools\private\initSize at 73
  In imshow at 262
  In megethos_leksis at 15
most common main body size is 29 pixels

ans =

    29
```

Fig.7 Sample output on image with 5 degrees vertical skew

References

[1] U.V. Marti, H. Bunke, “Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system”, Int. Journal of Pattern Recognition and Artifical Intelligence, 15(1), pp. 65–90, 2001.

[2] M. Côté, E. Lecolinet, M. Cheriet, C.Y.Suen, “ Automatic reading of cursive scripts using a reading model and perceptual concepts, The PERCEPTO system”, IJDAR, vol 1, pp. 3-17, 1998.

[3] C. K. Cheng and M. Blumenstein, “The neural-based segmentation of cursive words using enhanced heuristics”, Proc. of the 8th International Conference on Document Analysis and Recognition, pp.650-654, 2005.

[4] H. Lee and B. Verma, A novel multiple experts and fusion based segmentation algorithm for cursive handwriting recognition, Proc. of the International Joint Conference on Neural Networks, pp.2994-2999, 2008.

[5] A. Vinciarelli, J. Luettin, A new normalization technique for cursive handwritten words, Pattern Recognition Lett. 22 (9) (2001) 1043–1050.

[6] B. Gatos, I. Pratikakis, and K. Ntirogiannis. Segmentation based recovery of arbitrarily warped document images. In Proc. Int. Conf. on Document Analysis and Recognition, Curitiba, Brazil, Sep. 2007.

[7] D. Sharma, W. Shilpi, "Dewarping Machine Printed Documents of Gurmukhi Script", Information Systems for Indian Languages, Communications in Computer and Information Science series , v.139, pp. 117-123, 2011.

[8] N. Doulgeri, E. Kavallieratou, Retrieval of Historical Documents by Word Spotting, IS&T/SPIE Electronic Imaging 2009.

[9] T. Adamek, N. E. Connor, A. F. Smeaton, Word matching using single closed contours for indexing handwritten historical documents, Int. J. Doc. Anal.Recognit. 9 (2) (2007) 153–165.

[10] J. Rodriguez and F. Perronnin, "Handwritten word-spotting using hidden Markov models and universal vocabularies," Pattern Recognition, vol. 42, no. 9, pp. 2106–2116, 2009.

[11] E. Kavallieratou, N. Dromazou, N. Fakotakis, G. Kokkinakis, An Integrated System for Handwritten Document Image Processing, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 17, No. 4, pp. 101-120, 2003.

[12] Nikos Vasilopoulos, Ergina Kavallieratou, "A classification-free word-spotting system", Proc. SPIE 8658, Document Recognition and Retrieval XX, 2013.

[13] V. Papavassiliou, T. Stafylakis, V. Katsouros, G. Carayannis, "Handwritten Document Image Segmentation into Text Lines and Words, Pattern Recognition", Vol.43, No.1, pp. 369-377, 2010.

[14] http://www.unipen.org/trigraphslant.html

ELCVIA

Electronic Letters on Computer Vision and Image Analysis